Cyberbullying Detection with BiRNN and Attention Mechanism

Anman Zhang[1(✉)], Bohan Li[1,2,3], Shuo Wan[1], and Kai Wang[1]

[1] College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics, Nanjing, China
{zhanganman,bhli}@nuaa.edu.cn
[2] Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, China
[3] Jiangsu Easymap Geographic Information Technology Corp., Ltd., Nanjing, China

Abstract. While the social network has brought a lot of conveniences to our lives, it has also caused a series of severe problems, which include cyberbullying. Cyberbullying is an aggressive and intentional act carried out by a group or an individual to attack a victim on the Internet. Most of the existing works related to cyberbullying detection focus on making use of swearwords to classify text or images with short titles. Although previous methods such as SVM and logistic regression show some advantages in the accuracy of detection, few of them capture the semantic information of non-swearwords which could also make big difference to the final results. In this paper, we propose to use BiRNN and attention mechanism to identify bullies. BiRNN is used to integrate the contextual information, and the attention model reflects the weight of different words for classification. Meanwhile, we convert the severity calculated by the attention layer to the level of cyberbullying. Experiments conducted on three real-world text datasets show that our proposed method outperforms the state-of-art algorithms on text classification and identification effect.

Keywords: Attention model · Cyberbullying detection · Text classification · Social network

1 Introduction

An increasing number of people are suffering from cyberbullying in social network, especially adolescents. Cyberbullying is defined as "an aggressive and

This work is supported by National Natural Science Foundation of China (61728204, 61672284), Key laboratory of high security system software development and verification technology, ministry of industry and information technology (NJ2018014), Youth science and technology innovation fund (NT2018028), Foundation of state key laboratory of smart grid protection and operation control (national defense pre-research field), Jiangsu university advantageous discipline construction support project.

X. B. Zhai et al. (Eds.): MLICOM 2019, LNICST 294, pp. 623–635, 2019.
https://doi.org/10.1007/978-3-030-32388-2_52

intentional act carried out by a group or an individual repeatedly and continuously against victims through digital devices" [1]. In 2016, a study on 5,700 middle school and high school students aged 12–17 in the United States showed that 33.8% of them have long been cyberbullied [2]. Girls are more likely than boys to be disturbed by such problem. Meanwhile, another analysis on the prevalence of cyberbullying in Chinese college students [3] demonstrated that 39.18% of 781 subjects are involved in cyberbullying. Cyberbullying has been a widespread problem and could cause potential psychological harm to people.

Solving cyberbullying has received enormous attention in recent years. Some previous work based on dictionary matching counts the frequency of characteristic words as the evidence for text classification. Such methods depend on artificially designed features, so they cannot capture the contextual information, and the level of cyberbullying is difficult to be measured. Afterwards, some work starts to focus on text representation obtained by deep learning frameworks. Rosa et al. [4] reviewed related methods such as CNN, a hybrid CNN-LSTM and a mixed CNN-LSTM-DNN for cyberbullying detection. Cheng et al. [5] designed a hierarchical attention network to aggregate words into session vectors layer by layer.

Compared with them, we use BiRNN to identify various cyberbullying instances by analyzing social messages on the Internet, and then measure the severity of them. To summarize, we make the following contributions:

(1) We propose a novel model that incorporates attention into BiRNN to classify all the text into two types. One type contains cyberbullying content, while the other does not contain. Visualized attention values on the test set assist to select cyberbullying topics with high attention value for detection.
(2) According to weights of the attention layer, we assume that the influence of other roles like defenders is negligible, and mainly identify bullies who play the leading role in the cyberbullying event.
(3) We measure the severity of cyberbullying with attention values on three datasets. Experimental results demonstrate the identification accuracy and the severity level.

Table 1 lists all the notations and descriptions. The rest of this paper is organized as follows. Section 2 briefly overviews the related work. Section 3 introduces the two stages involved in the research process. The first stage focuses on text classification with BiRNN and Attention Model, and in the second stage we identify bullies and measure the degree of cyberbullying. Experiments are carried out in Sect. 4 to verify the effectiveness of our method. Conclusion is drawn in Sect. 5.

2 Related Work

Most of the existing methods related to cyberbullying detection could be divided into four categories: content-based, sentiment-based, user-based and network-based.

Table 1. Notations and descriptions.

Notations	Descriptions
S_i	The sentence i
w_{in}	The word vector n of the sentence i
W	A word vector matrix
$\overrightarrow{h_{in}}$	The hidden state in the bidirectional recurrent neural network corresponding to the word vector w_{in}
s_{in}	A scoring function based on the degree of correlation between $\overrightarrow{h_{in}}$ and final types
W_a	The weight of attention mechanism
b_a	The bias of attention mechanism
a_{in}	The attention value corresponding to the word vector w_{in}
C	The context information
b_{att}	The average attention value of the bully
$asst_{i,att}$	The average attention value of the assistant i
p_b	The number of posts written by the bully
p_{asst_i}	The number of posts written by the assistant i

Content-Based. Rafiq et al. [6] argued that profanity is not the only feature for cyberbullying detection. The classifier should be supplemented by other indicators such as the profile of user, media session, and comment features. Then Nahar et al. [7] took other features such as pronouns into consideration for further cyberbullying detection. Specially, on the MySpace platform, Dadvar et al. [8] classified the corpus by gender and trained SVM as the classifier. TFIDF is a way to measure the frequency of foul words used by men and women. Another content-driven detection [9] explored the relationship between text and visual content concerning cyberbullying. Methods in this paper are related to deep learning and unsupervised clustering.

Sentiment-Based. Sentiment analysis is closely related to pronoun usage and TFIDF. Nahar et al. [10] used Probabilistic Latent Semantic Analysis (PLSA) to analyze labeled bullying posts which include potential sentiment features, then the most influential people, i.e., victims and predators are detected and ranked. Xu et al. [11] leveraged Twitter Streaming API to identify "bully traces" and concluded eight roles played by people referenced within the tweets, which include bully, victim, bystander, assistant, defender, reporter, accuser, and reinforcer. In their follow-up work, they tried to find emotions expressed in tweets by training a SVM classifier using distant labeled data from Wikipedia.

User-Based. Compared with content-based and sentiment-based features, user-based features tend to be ignored. Recently, some efforts have been made to add user-related features into cyberbullying systems, such as gender, age, and race. Nahar et al. [12] used a multi-agent system to deal with streaming data from multiple social network sources. Under the circumstance of insufficient labelled data, they could still detect cyberbullying automatically. Chen et al. [13] observed users' conversation history and writing styles in order to form Lexical Syntactic Framework (LSF) and gave an offensiveness score for users.

Network-Based. Within the field of cyberbullying detection, researchers also pay more attention to network data such as number of friends, uploads and likes. To improve the effect of detection, Dadvar et al. [14,15] used number of uploads, membership duration, comments, and subscriptions as features, while the ego network was used by NaliniPriya and Asswini [16] to obtain temporal changes in the relationship among users, which are valuable in the process of detection. Online social network topology structure was referred in the paper of Chelmis et al. [17].

Various techniques have been applied to identify cyberbullying incidents automatically based on those mentioned features above, and most of them adopt supervised learning techniques which were first used by Yin et al. [18]. Potha and Maragoudakis [19] modeled data via three feature representation formats: BoW, weights allocation with SVM, and feature space simplification with SVD. Squicciarini et al. [20] used a C4.5 Decision Tree classifier based on content, personal features, and social network features to identify bullies.

However, deep learning methods are rarely mentioned although it has gained popularity in recent years. In contrast, our method is a combination of content-based features and sentiment-based features. It is superior to traditional methods in accuracy and displays the detection effect visually. Besides, bullies can be detected, and the cyberbullying degree can be measured.

3 Attention Detection Model

Attention is intuitively how much people pay attention to what they are interested in. The aim of constructing the attention detection model is summarized as two points. One is to train the classification model, the other is to measure the cyberbullying. The core part of the proposed method is the attention layer, where we calculate the average attention of every post and user to identify and measure the bullies.

3.1 Problem Description

Definition 1 *(Attention Detection). Assume that there is a topic T under which m users $(u_1, u_2, \ldots, u_m)$ send posts. Users in this topic include the bully, assistants and others whose posts are not related to the cyberbullying. The posts sent by each user are regarded as $post_i$, $i = 1, 2, \ldots, n$. Attention detection is aimed at measuring different effect of each word in $post_i$ of m users for classification, thus determining whether the cyberbullying is occurring.*

As shown in Fig. 1, the research framework is divided into two stages. The first stage is meant to classify all the text into two types, i.e. positive and negative.

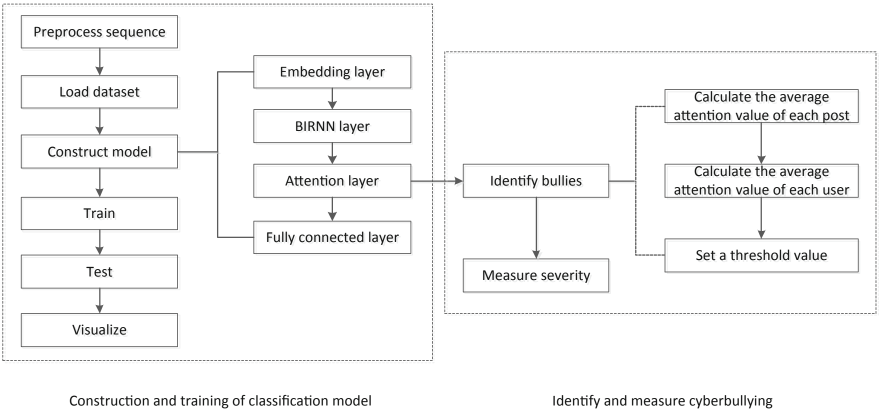

Fig. 1. The research framework.

The negative type includes cyberbullying information, while the positive type contains non-cyberbullying information or few statements defending the victim. The second stage is meant to identify bullies who make offensive remarks about others and measure the severity of cyberbullying.

3.2 Attention Detection Based on BiRNN

During the process of constructing classification model, we use a model which combines Bidirectional Recurrent Neural Network (BiRNN) with Attention Model (AM). The BiRNN is designed to bind RNN that moves from the beginning of the sequence and RNN that starts at the end of the sequence, namely the forward layer and the backward layer respectively. It tends to average the accumulated output vectors of each time, assuming that each input word contributes equally to the text representation. While some words should own much greater weights, in particular sensitive words. Therefore, we incorporate attention model into BiRNN to classify text.

To illustrate it, Fig. 2 shows part of conversation in one topic and the symbol '+' separates different posts. The shades of color vary due to the difference in the attention value assigned to each word. It is understandable that the color of some signal words like 'stalker', 'faggotry' and 'homosexual' is much darker. However, the color of other non-swearwords also changes more or less. If these words are extracted separately, they seem to have nothing to do with cyberbullying. When we put them together in the context, their meanings are changed.

3.3 Text Classification

The process of text classification is the prerequisite to identify bullies and calculate the degree of cyberbullying. As shown in Fig. 3, it includes four steps:

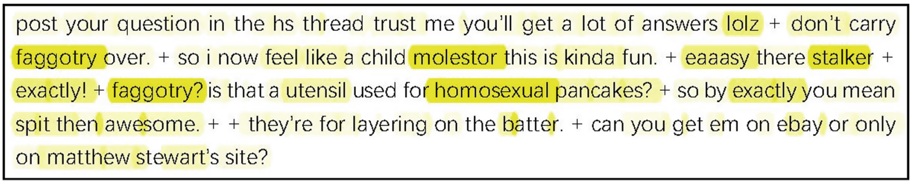

Fig. 2. One example concerning cyberbullying. (Color figure online)

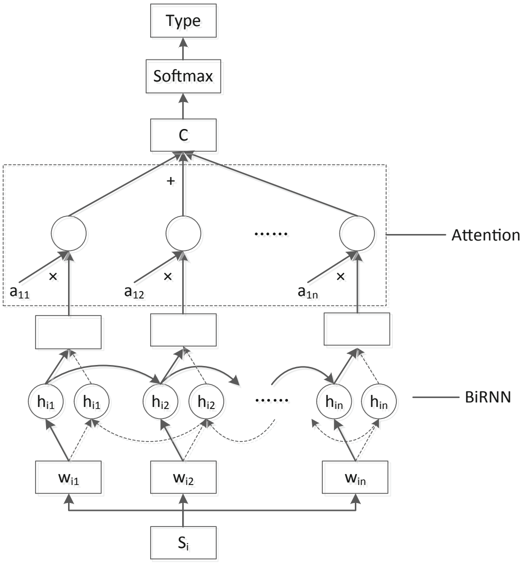

Fig. 3. Text classification process with BiRNN and attention model.

(1) The sentence S_i is divided into words, and each word is translated into the word vector sequence $w_{i1}, w_{i2}, \ldots, w_{in}$. Each sentence corresponds to a matrix $W = (w_{i1}, w_{i2}, \ldots, w_{in})$.
(2) In the BiRNN layer, each word is seen as a time node. The word vector is the input feature of each unit. The output vector of each time step is regarded as a contribution made by the corresponding input to the current task in the context. The forward and backward layers are composed of bidirectional characteristics. Each unit has state characteristics $\overrightarrow{h_{in}}$ and output characteristics after calculation. Output characteristics of each unit are used to calculate the weighted average of the attention layer.
(3) Formula (1) is a scoring function based on the degree of correlation between $\overrightarrow{h_{in}}$ and the final type. The more relevant, the greater s_{in} is. Here W_a and b_a are the weight and bias of attention.

$$s_{in} = tanh\left(W_a\overrightarrow{h_{in}} + b_a\right) \tag{1}$$

For all s_{in}, the final attention value a_{in} is obtained by a softmax function. The word context vector u_w is initialized randomly and learned jointly during the training process.

$$a_{in} = \frac{e^{s_{in}^T u_w}}{\sum_n e^{s_{ik}^T u_w}} \tag{2}$$

The context information $c(1 \leq x \leq)$ for text classification is represented as Formula (3).

$$\begin{cases} h_{i1} * a_{11} + h_{i2} * a_{12} + \ldots + h_{in} * a_{1n} = c_1, \\ h_{i1} * a_{21} + h_{i2} * a_{22} + \ldots + h_{in} * a_{2n} = c_2 \end{cases} \tag{3}$$

(4) We use the fully connected layer and the softmax function to output the probability of each type. Whether the text belongs to cyberbullying is dependent on these probabilities.

3.4 Cyberbullying Identification and Measurement

The second stage is to detect bullies who send aggressive posts and evaluate the severity of cyberbullying. Rather than count the frequency of swearwords, we compute the average attention value of all users in each topic containing cyberbullying information to set a threshold. As shown in Algorithm 1, threshold reflects the average level of attack.

Among all users whose average weight is above the threshold, the user with the highest value is the dominant bully, and the others can be regarded as assistants promoting this terrible event. Algorithm 2 illustrates the process of identifying the bully and assistants. Then, to calculate the severity of each topic that contains the bullying content, i.e. the level of attack, we regard attention values b_{att} and $asst_{i,att}$ pointing to cyberbullying type as the degree of attack. The number of posts written by per person is considered as a weight.

$$severity = \frac{b_{att} \times p_b + \sum (asst_{i,att} \times p_{asst_i})}{p_b + \sum p_{asst_i}} \tag{4}$$

In Formula (4), b and $asst_i$ represent the bully and assistants in one topic. There is only one bully and there are many assistants. p_b and p_{asst_i} are the number of posts from the bully and each assistant.

4 Experiment

4.1 Dataset Collection

We conduct experiments on three datasets of the social network: Formspring, Twitter, and MySpace. Formspring is a question and answer platform launched in 2009. Twitter provides the microblogging service that allows users to update messages within 140 characters. MySpace is a social website, which provides

Algorithm 1. Threshold setting in the cyberbullying topic

Input: the post j of the user i, $post[i][j]$; the attention value j of the user i, $attention[i][j]$; the number of the bully and assistants, N; the attention value of the word, a

Output: threshold

```
p_b ← 0, p_asst_i ← 0, i ← 0, j ← 0, count ← 0, sum ← 0
for i < N do
    for word in post[i][j] do
        sum ← sum + a
        count++
    end for
    attention[i][j] ← sum/count
    i++
end for

i ← 0, j ← 0, count ← 0, sum ← 0
for i < N do
    if attention[i][j] exists then
        sum ← sum + attention[i][j]
        count++
        j++
    else
        ave[i] ← sum/count
        i++, j ← 0, count ← 0, sum ← 0
    end if
end for

i ← 0, sum ← 0
for i < N do
    sum ← sum + ave[i]
    threshold ← sum/N
end for
return threshold
```

global users with an interactive platform integrating social networking, personal information sharing, instant messaging, and other functions.

Formspring[1]. This dataset contains 40,952 posts from 50 ids in Formspring. Each post is crowdsourced to three workers of Amazon Mechanical Turk (AMT) for labeling the cyberbullying content with ‘Yes’ or ‘No’. About 3,469 posts are regarded as the bullying type by at least one worker and 37,349 posts are deemed non-cyberbullying. The rest of the data is not given a definitive judgment.

Twitter[2]. This dataset is collected from Twitter stream API. It has 7,321 tweets consisting of 2,102 ‘y’ posts and 5,219 ‘n’ posts. All the data has been labeled by experienced annotators for cyberbullying research.

[1] http://www.chatcoder.com/drupal/DataDownload.
[2] http://research.cs.wisc.edu/bullying/data.html.

Algorithm 2. Identify the bully and assistants

Input: N, $ave[i]$, p_b, p_{asst_i}, threshold
Output: b_{att}, $asst_{t,att}$

$i \leftarrow 0$, $j \leftarrow i+1$
$max \leftarrow ave[i]$
for $j < N$ **do**
  if $max < ave[j]$ **then**
    $max \leftarrow ave[j]$
  end if
  $j{+}{+}$
end for

$b_{att} \leftarrow max$
for $i < N$ **do**
  if $ave[i] == max$ **then**
    continue
  else if $ave[i] > threshold$ **then**
    $asst_{i,att} \leftarrow ave[i]$
  end if
  $i{+}{+}$
end for
return b_{att}, $asst_{i,att}$

MySpace (See footnote 1). There are 381,557 posts that belong to 16,345 topics in the extended dataset. Firstly, we save swear words, bad words and curse words from a website called Swear Word List & Curse Filter[3]. Some Internet slang[4] and British slang[5] consisting of slang and acronyms that include foul words are also selected. Then we match these words with contents of all the posts to label each post automatically. If a post contains the bullying content, it is labelled as '1', otherwise it is labelled as '0'. Among all the topics, there are 10,629 labels for '1' and 5,716 labels for '0'.

4.2 Results and Analyses

In the process of training, we vary the learning rate to compare experimental results and seek the best parameter. Figure 4 demonstrates the parameter tuning on MySpace dataset. The x-coordinate represents iteration times, and the y-coordinate represents the accuracy and the cross entropy loss respectively. The overall trend of accuracy is on the rise and the loss is declining. It is obvious that the accuracy and loss reach a balance when the learning rate is set to 1e−3. Parameter adjustments on other datasets are similar.

To verify the effectiveness of the proposed model on the test set, we set the iteration times of the experiment to 20 and repeat the experiment for 5 times.

[3] https://www.noswearing.com/dictionary.
[4] https://www.noslang.com/dictionary.
[5] https://www.translatebritish.com/dictionary.

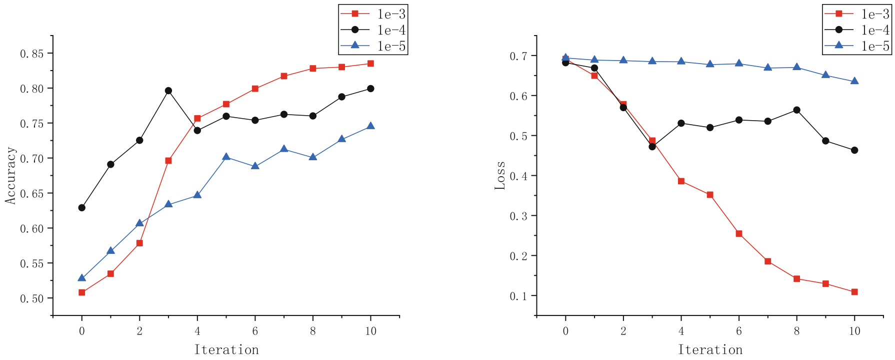

Fig. 4. Changes in accuracy and loss of test set varying the learning rate.

The average value of all indicators are taken as the final results. We take SVM, Logistic Regression, and CNN referred before as baseline algorithms. The effectiveness of these algorithms is evaluated from three aspects: accuracy, precision, and recall. As shown in Fig. 5, our method outperforms benchmark methods on all datasets. The traditional classification methods of artificial feature extraction such as SVM and Logistic Regression are inferior to the deep learning like CNN. Although our model is slightly better than CNN, we exploit the attention mechanism to detect the cyberbullying.

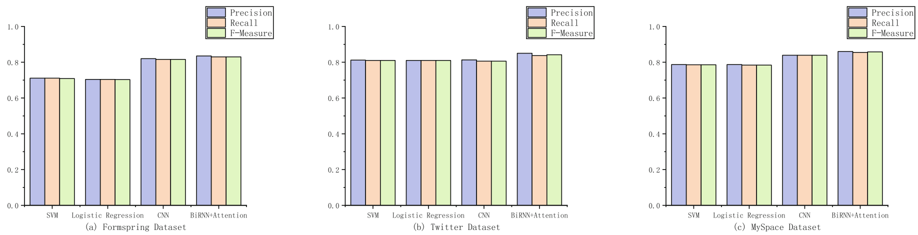

Fig. 5. Comparison between BiRNN+Attention and three advanced algorithms.

Taking MySpace as an example, Fig. 6 visualizes the impact of each word on the final classification. This topic is about ‘Weed or Alcohol’ and it contains 50 posts. The number of all words in these posts are nearly 1000. We mark several words with high weights in the following figure. Some insulting words like ‘fuck’, ‘stalker’ and ‘faggotry’ have higher weights, while some emotional words like ‘lol’ also deserve to be paid attention to. That is why we introduce attention mechanism into the process of identification, instead of depending on the number of swearwords. According to the distribution of attention value, the scope of identifying the bully can be reduced. Text with high and dense attention values calls for the special attention. It is most likely to be posted by the bully.

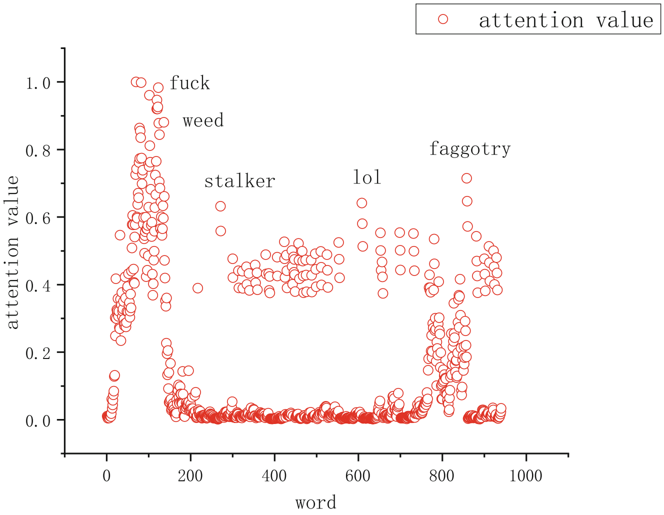

Fig. 6. The attention value of all words in one cyberbullying topic

With a preliminary analysis of attention values, we try to identify the bully and assistants. The first thing that we need to do is determining the threshold. We compute the average attention value of all users who belong to files marked as cyberbullying. Next, by Formula (4), the severity of cyberbullying can be measured. Table 2 shows the results of bully detection and severity measure. We select 1,000 posts labeled as cyberbullying from three test sets respectively. Finally, we convert the probabilities of the severity into the corresponding level from 1 to 9. It can be seen that the bullying problem in Twitter is more serious than other two platforms. It makes sense because Twitter is intuitively the most popular social platform of them. In addition, the more posts a bully sends, the worse the problem tends to be.

Table 2. Bully detection and severity measure.

Dataset	Posts	Bully	p_b	Threshold	Severity	Level
Formspring	1,000	joie***esu	22	0.48	0.5715	5
Twitter	1,000	31***104	56	0.66	0.7887	7
MySpace	1,000	$MS_1$3***63	8	0.23	0.3119	3

5 Conclusion and Future Work

In this paper, we study the problem of cyberbullying text detection and severity measure. We propose a model that takes advantage of BiRNN and Attention

Model to classify text. Weights of the attention layer are further used to identify the bully and measure the severity of cyberbullying. Visualized attention value assists in reducing the detection range of bullying text, and remarkable words are expected to be saved for building a knowledge base. Experimental results on three different datasets show that the proposed model is effective. Considering the many-to-many relationship between bully and victim, we will try to find the cyberbullying community instead of one dominated bully in our future work.

References

1. Smith, P.K.: Cyberbullying: its nature and impact in secondary school pupils. J. Child Psychol. Psychiatry **49**(4), 376–385 (2008)
2. Cyberbullying Data. https://cyberbullying.org/2016-cyberbullying-data. Accessed 26 Nov 2016
3. Zhu, H.: Analysis on prevalence of cyberbullying in college students in China. J. Jilin Univ. (Med. Ed.) **42**(3), 605–611 (2016)
4. Rosa, H., Matos, D., Ribeiro, R., Coheur, L., Carvalho, J.P.: A "Deeper" look at detecting cyberbullying in social networks. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE, Rio (2018)
5. Cheng, L., Guo, R., Silva, Y., Hall, D., Liu, H.: Hierarchical attention networks for cyberbullying detection on the Instagram social network. In: SIAM International Conference on Data Mining. SDM, Calgary (2019)
6. Rafiq, R.I., Hosseinmardi, H., Han, R., Lv, Q., Mishra, S., Mattson, S.A.: Careful what you share in six seconds: detecting cyberbullying instances in Vine. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 617–622. ACM, Paris (2015)
7. Nahar, V.: An effective approach for cyberbullying detection. Commun. Inf. Sci. Manag. Eng. **3**(5), 238 (2013)
8. Dadvar, M., De Jong, F.: Cyberbullying detection: a step toward a safer internet yard. In: Proceedings of the 21st International Conference on World Wide Web, pp. 121–126. ACM, Lyon (2012)
9. Zhong, H., et al.: Content-driven detection of cyberbullying on the Instagram social network. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, pp. 3952–3958. AAAI, New York (2016)
10. Nahar, V., Unankard, S., Li, X., Pang, C.: Sentiment analysis for effective detection of cyber bullying. In: Sheng, Q.Z., Wang, G., Jensen, C.S., Xu, G. (eds.) APWeb 2012. LNCS, vol. 7235, pp. 767–774. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29253-8_75
11. Xu, J.M., Jun, K.S., Zhu, X., Bellmore, A.: Learning from bullying traces in social media. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 656–666. Association for Computational Linguistics, Montreal (2012)
12. Nahar, V.: Detecting cyberbullying in social networks using multi-agent system. Web Intell. Agent Syst.: Int. J. **12**(4), 375–388 (2014)
13. Chen, Y., Zhou, Y., Zhu, S., Xu, H.: Detecting offensive language in social media to protect adolescent online safety. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, pp. 71–80. IEEE, Amsterdam (2012)

14. Dadvar, M., Trieschnigg, R.B., de Jong, F.M.: Expert knowledge for automatic detection of bullies in social networks. In: 25th Benelux Conference on Artificial Intelligence, BNAIC 2013, pp. 57–64. Delft University of Technology, TU Delft (2013)
15. Dadvar, M., Trieschnigg, D., de Jong, F.: Experts and machines against bullies: a hybrid approach to detect cyberbullies. In: Sokolova, M., van Beek, P. (eds.) AI 2014. LNCS, vol. 8436, pp. 275–281. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-06483-3_25
16. Salawu, S.: Approaches to automated detection of cyberbullying: a survey. IEEE Trans. Affect. Comput. (2017). https://doi.org/10.1109/TAFFC.2017.2761757
17. Chelmis, C., Zois, D.S., Yao, M.: Mining patterns of cyberbullying on Twitter. In: 2017 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 126–133. IEEE, New Orleans (2017)
18. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., Edwards, L.: Detection of harassment on Web 2.0. In: Proceedings of the Content Analysis in the WEB, vol. 2, pp. 1–7. CAW2.0, Madrid (2009)
19. Potha, N., Maragoudakis, M.: Cyberbullying detection using time series modeling. In: 2014 IEEE International Conference on Data Mining Workshop, pp. 373–382. IEEE, Shenzhen (2014)
20. Squicciarini, A., Rajtmajer, S., Liu, Y., Griffin, C.: Identification and characterization of cyberbullying dynamics in an online social network. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 280–285. ACM, Paris (2015)